

**Московский авиационный институт
(Национальный исследовательский университет)**

Факультет: «Информационные технологии и прикладная математика»

Кафедра: 806 «Вычислительная математика и программирование»

Дисциплина: «Машинное обучение»

Лабораторная работа № 1
Тема: Azure Machine Learning.

Студент: Мамчур А.В.

Группа: 80-304

Дата:

Постановка задачи

Ваша задача познакомиться с платформой Azure Machine Learning, реализовывая полный цикл разработки решения задачи машинного обучения, используя три различных алгоритма, реализованные на этой платформе.

Требования

- Уникальность решения
- Обоснованность выбора той или иной операции
- В отчете должны быть указаны алгоритмы, которые применялись, результаты применения этих алгоритмов, а также скрины некоторых этапов обработки данных

Ход работы

Датасет: <https://www.kaggle.com/donpiano/motorcycle-listings> (мотоциклы)

1) Алгоритм линейной регрессии(Linear regression)

Прогнозируемая цель: цена на мотоцикл.

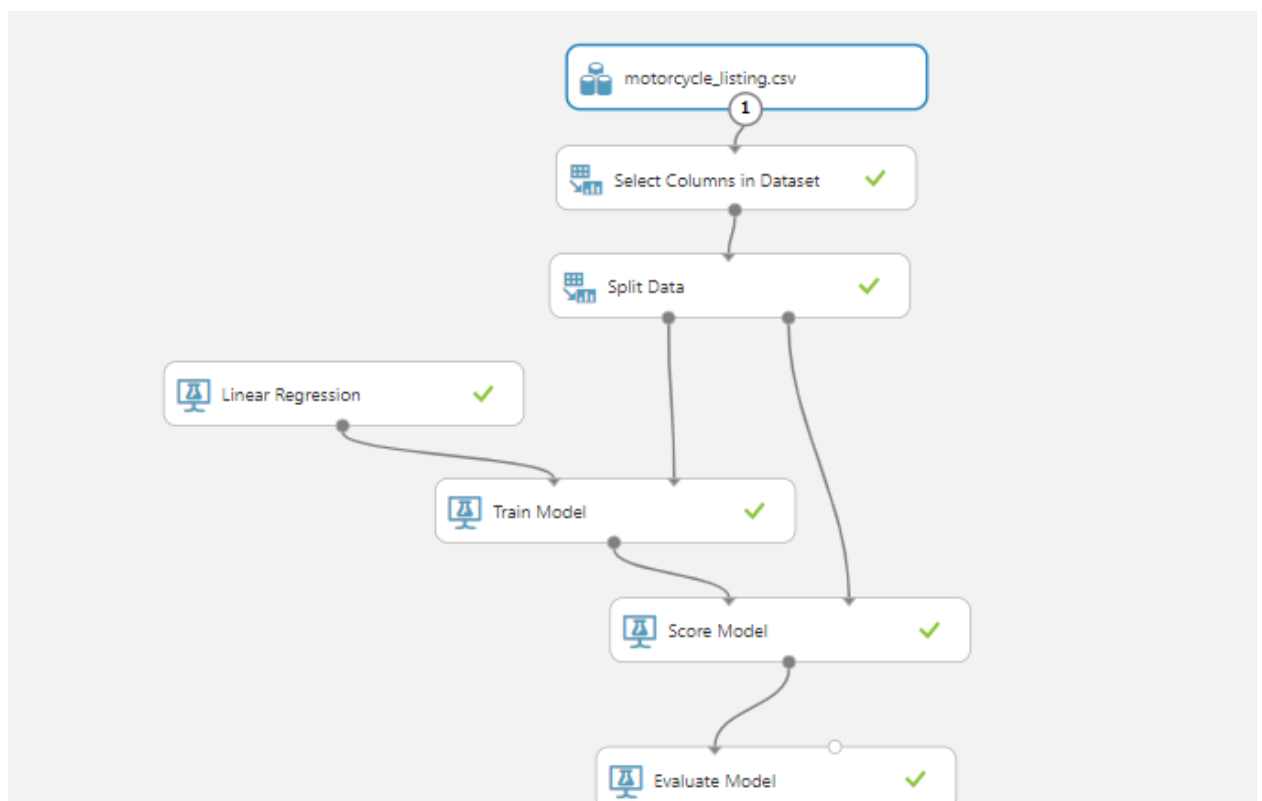
Подготовка данных

Некоторые свойства лучше подходят для прогнозирования данной цели, чем другие.









Поэтому с помощью модуля *Select Columns in Dataset* (Выбор столбцов в наборе данных) можно убрать некоторые ненужные столбцы(unique_id, number_images). Эти столбцы не имеют отношения к цене.

Модуль *Split Data* позволяет разделить набор данных на два, один из которых применим для обучения модели(75%), а второй — для тестирования(25%).

Модуль линейной регрессии использует метод наименьших квадратов для проецирования линии, проходящей через все точки данных, присутствующие в наборе данных обучения.



На порту вывода модуля *Score Model* будут показаны прогнозируемые значения цены (Scored Labels) вместе с известными значениями проверочных данных (price):

	brand	model	name	year	price	body_type	Scored Labels
view as 							
	Harley-Davidson	EL Knucklehead	1941 Harley-Davidson EL Knucklehead	1941	35000	Vintage	46540.695775
	KTM	85 SX (Big Wheel)	2015 KTM 85 SX (Big Wheel)	2015	3999	Motocross 2 Stroke	3500.91636
	Yamaha	YZ125	2018 Yamaha YZ125 MY19	2018	8990	Motocross 2 Stroke	5145.680283
	Triumph	Steve McQueen SE	2012 Triumph Steve McQueen SE	2012	20700	Naked	10316.036782
	BMW	S 1000 RR	2016 BMW S 1000 RR	2016	18490	Super Sport	19290.832433
	Ducati	916	1996 Ducati 916	1996	17000	Super Sport	19294.042258
	Chrome	Flash	2007 Chrome Flash	2007	26000	Cruiser	19921.476274

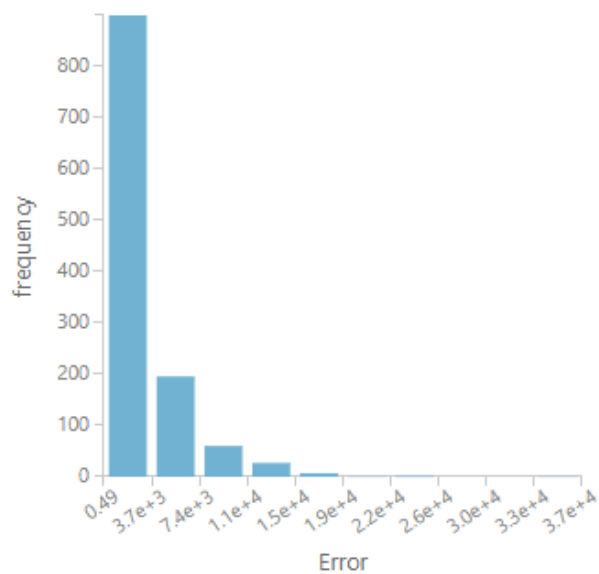
Также мы можем проверить качество результатов с помощью модуля *Evaluate Model*. Для нашей модели будет выведена следующая статистика.

- **Средняя абсолютная погрешность.** Среднее значение абсолютной погрешности (*погрешность* — это разница между спрогнозированным и фактическим значением).
- **Среднеквадратичное отклонение.** Квадратный корень из среднего значения возведенных в квадрат арифметических отклонений спрогнозированных значений тестового набора данных.
- **Относительное арифметическое отклонение.** Среднее арифметическое отклонение по отношению к абсолютной разнице между фактическими значениями и средним арифметическим всех фактических значений.
- **Относительное среднеквадратичное отклонение.** Среднее арифметическое среднеквадратичных отклонений по отношению к абсолютной разнице между фактическими значениями и средним арифметическим всех фактических значений.
- **Коэффициент смешанной корреляции (R в квадрате).** Статистический показатель, который оценивает соответствие модели данным.

Metrics

Mean Absolute Error	2761.714049
Root Mean Squared Error	4299.717273
Relative Absolute Error	0.424891
Relative Squared Error	0.229336
Coefficient of Determination	0.770664

Error Histogram

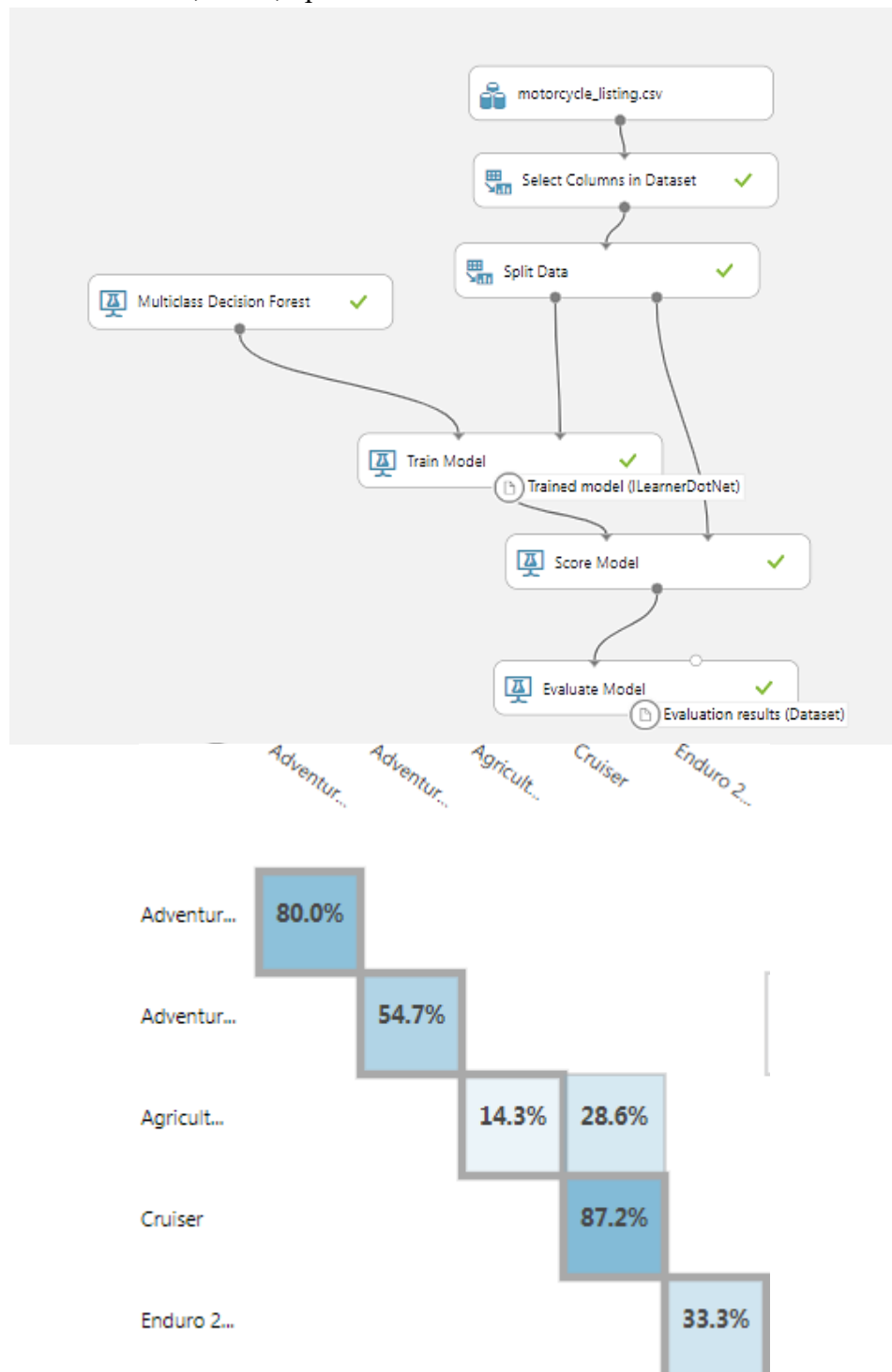


Как показано выше, наша модель линейной регрессии имеет значение R-квадрата (коэффициент определения) 0,77 . Это означает, что 77% отклонения в ценовых значениях может быть описано изменением значений столбцов признаков (предикторов).

2) Алгоритм мультиклассовой классификации (Multiclass Decision Forest)

Задача: классификация мотоциклов по типам.

Выбранные поля: модель, цена, бренд.



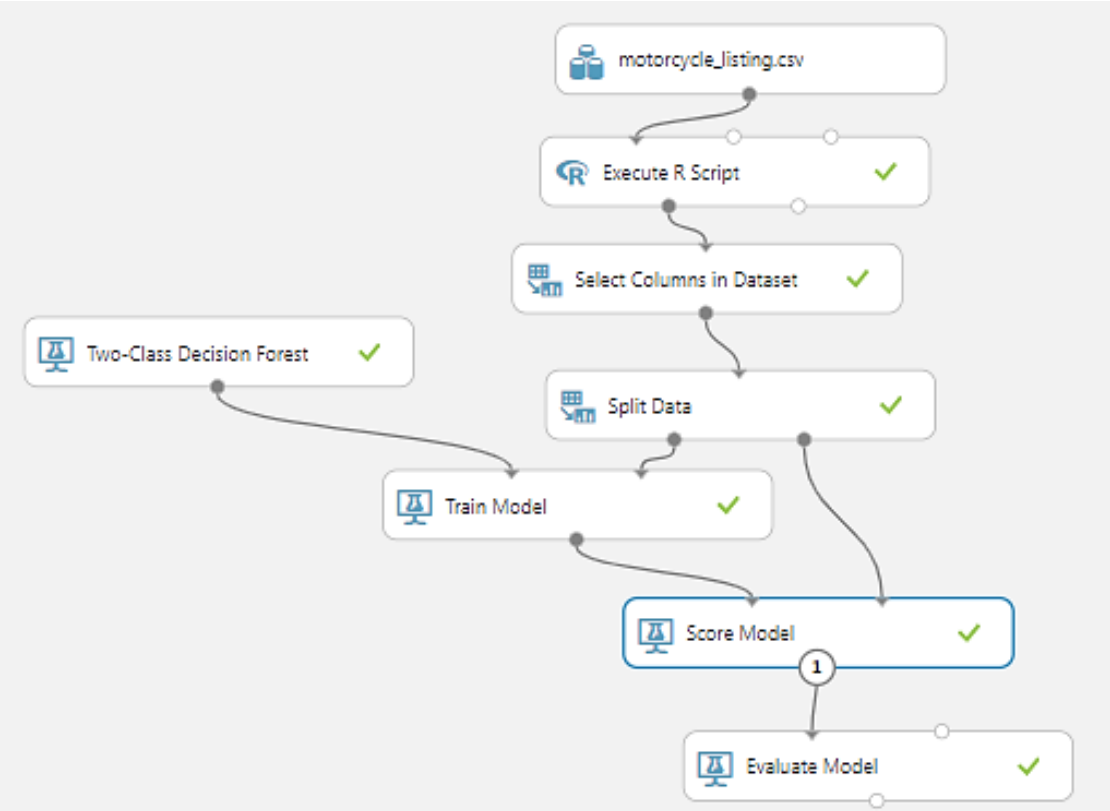
Матрица неточностей

Матрица неточностей используется для описания эффективности модели классификации. Каждая строка сопоставлена с экземпляром реального класса, а каждый столбец — с экземпляром прогнозируемого класса. Матрица неточностей позволяет увидеть количество правильно классифицированных выборок из тестового набора.

Как видно из полученной матрицы, большинство типов классификатор определяет верно. Диагональные элементы матрицы явно выражены. Тем не менее в рамках некоторых классов (14, 33) классификатор показывает низкую точность.

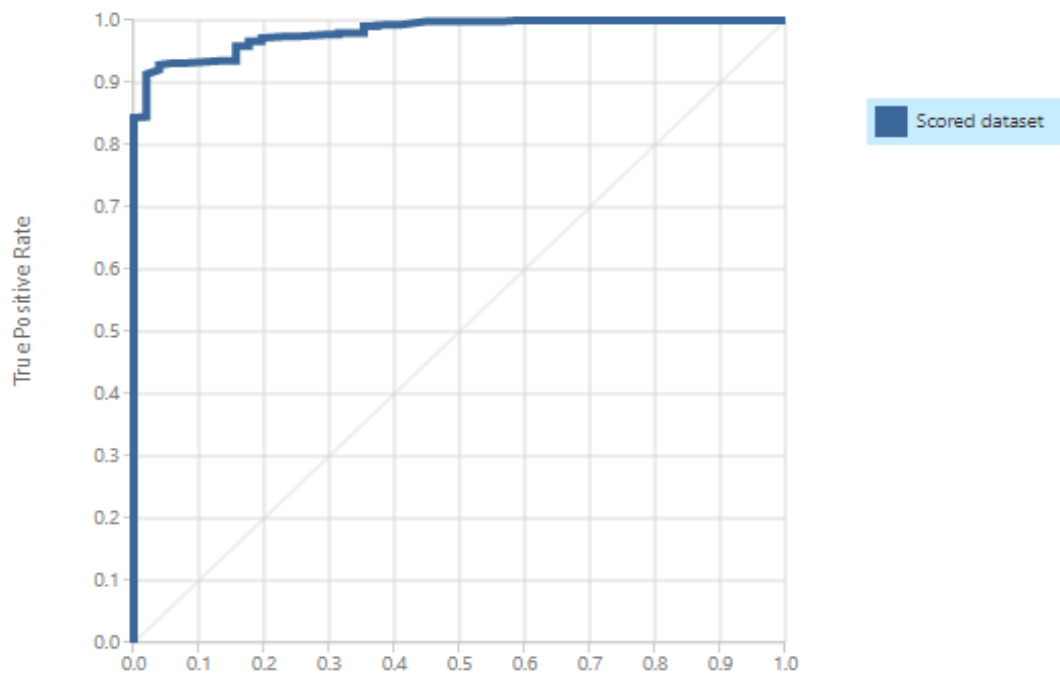
3) Алгоритм двухклассовой классификации(Two-Class Decision Forest)

Чтобы воспользоваться алгоритмом Two-Class Decision Forest, я добавила в свой датасет с помощью скрипта на R дополнительный столбец с бинарными значениями и провела классификацию мотоциклов по году выпуска(ранее 2000 -1 , позже 2000 - 0).



name	year	price	body_type	km	vehicle_engine	result	Scored Labels	Scored Probabilities
2007 Honda VTR250	2007	2200	Naked	33850	248	1	1	0.873004
2007 Yamaha TW200E	2007	4600	Trail	6772	196	1	1	0.95392
2008 Ducati 1098 R	2008	30000	Super Sport	10320	1198	1	1	0.969803
2007 Suzuki Boulevard M109R (VZR1800)	2007	7500	Cruiser	42000	1783	1	1	0.920075

ROC PRECISION/RECALL LIFT



True Positive	False Negative	Accuracy	Precision	Threshold	AUC
742	0	0.938	0.938	0.5	0.981
False Positive	True Negative	Recall	F1 Score		
49	2	1.000	0.968		
Positive Label	Negative Label				
1	0				

Score Bin	Positive Examples	Negative Examples	Fraction Above Threshold	Accuracy	F1 Score	Precision	Recall	Negative Precision	Negative Recall	Cumulative AUC
(0.900,1.000]	688	2	0.870	0.929	0.961	0.997	0.927	0.476	0.961	0.035
(0.800,0.900]	53	22	0.965	0.968	0.983	0.969	0.999	0.964	0.529	0.452
(0.700,0.800]	1	13	0.982	0.953	0.976	0.953	1.000	1.000	0.275	0.707
(0.600,0.700]	0	7	0.991	0.945	0.971	0.944	1.000	1.000	0.137	0.844
(0.500,0.600]	0	5	0.997	0.938	0.968	0.938	1.000	1.000	0.039	0.942
(0.400,0.500]	0	2	1.000	0.936	0.967	0.936	1.000	1.000	0.000	0.981
(0.300,0.400]	0	0	1.000	0.936	0.967	0.936	1.000	1.000	0.000	0.981
(0.200,0.300]	0	0	1.000	0.936	0.967	0.936	1.000	1.000	0.000	0.981
(0.100,0.200]	0	0	1.000	0.936	0.967	0.936	1.000	1.000	0.000	0.981
(0.000,0.100]	0	0	1.000	0.936	0.967	0.936	1.000	1.000	0.000	0.981

Метрики качества классификации:

- *Accuracy* (аккуратность) — процент верных предсказаний
- *Precision* (точность) — сколько верных среди предсказанных как “Класс 1/да”.
 $Precision = TP / (TP + FP)$
- *Recall* (полнота) — сколько из настоящих “Класс 1/да” мы определили верно (то же самое, что True Positive Rate, Sensitivity)
 $Recall = TP / (TP + FN)$
- *F-мера* - гармоническое среднее точности и полноты.

ROC-кривая показывает зависимость количества верно классифицированных

положительных примеров от количества неверно классифицированных отрицательных примеров. Количественную интерпретацию ROC даёт показатель - площадь, ограниченная ROC-кривой и осью доли ложных положительных классификаций. Чем выше *показатель AUC*, тем качественнее классификатор. В моем случае $AUC = 0.981$, что близко к единице, значит классификатор хорошо справляется с задачей.

Вывод: Данная лабораторная работа позволяет познакомиться со студией машинного обучения Azure, которая значительно упрощает и ускоряет создание ML-систем, а также делает их более результативными. С помощью данного ресурса стало возможным проверить эффективность разных алгоритмов при решении одной и той же задачи без навыков программирования.